Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Claudius Korzen

Information Retrieval
WS 2017/2018

`http://ad-wiki.informatik.uni-freiburg.de/teaching`

Exercise Sheet 9

Submit until Tuesday, January 9 **2018 (next year)** at **12:00pm (noon)**

Exercise 1 (5 points)

Let A be an arbitrary $m \times n$ term-document matrix. Let μ_1 and μ_2 be two $m \times 1$ vectors which are L_2-normalized, that is, the sum of the squares of their entries is 1. Let M be the $m \times 2$ matrix consisting of μ_1 in the first column and μ_2 in the second column. Let D be the matrix product $M^T \cdot A$ and note that D is a $2 \times n$ matrix.

Then prove the following for each column (document) A_i of A: Let A'_i be the L_2-normalized version of A_i. Then A'_i is closer to μ_1 than to μ_2 with respect to the Euclidean distance if and only if D_{1i} is larger than D_{2i}, where D_{ij} is the entry of D at position i, j.

Exercise 2 (10 points)

Execute the k-means algorithm on the following term-document matrix A, using Euclidean distance between normalized documents as distance measure. The number of clusters is $k = 2$, and the initial centroids are the first and third column of A. When a document has the same distance to both centroids, assign it to the second centroid.

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

For each step of k-means, write down the centroid vectors and the clusters, and how you computed the assignments from the documents to the clusters. In order to compute that assignment in each step, you should use the result from Exercise 1. You must *not* use a calculator but write down the precise mathematical expressions (for example, $1/2 \cdot \sqrt{2}$). Avoid unnecessarily complicated calculations: if your expressions become convoluted, you are doing something wrong.

[turn over after convergence]

Exercise 3 (5 points)

Now consider a more general $m \times n$ term-document matrix of the following form, where A_1 is an $m_1 \times n_1$ matrix and A_2 is an $m_2 \times n_2$ matrix and $m_1 + m_2 = m$ and $n_1 + n_2 = n$ and no two columns of A_1 and no two columns of A_2 are exactly orthogonal to each other (that is, each pair of documents from the same "side" of A have at least one word in common):

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

Note that the two 0s are whole matrices filled with 0s. Let μ_1 be one of the first n_1 columns of A, and let μ_2 be one of the last n_2 columns of A. Then prove that if one starts k-means with centroids μ_1 and μ_2, the resulting clustering will not change anymore after the first step.

Add your submission *as a single PDF* to a new sub-directory *sheet-09* of your folder in the course SVN, and commit it. You may upload a handwritten solution, but only if it is neatly written and properly scanned. In all other cases, you should use LaTeX (other typesetting programs are not allowed). Also commit the usual *experiences.txt* with your brief and concise feedback.